Construction of Malaria Gene Expression Network Using Partial Correlations

Khanin, Raya; Wit, Ernst

Published in:
Methods of Microarray Data Analysis V

Publication date:
2007

Chapter 6

Construction of Malaria Gene Expression Network Using Partial Correlations

Raya Khanin and Ernst Wit

Department of Statistics, University of Glasgow, Glasgow, UK
e-mail: raya@stats.gla.ac.uk

Abstract In this paper we model the gene expression network of *Plasmodium falciparum* using the time-course microarray dataset [Bozdech, Z., et al., *PLoS Biol.*, **1**(1) (2003), E5] A gene expression network is constructed based on a novel method that combines two types of correlations between each pair of genes: standard Pearson and partial correlations. A link is established between two genes if both correlation coefficients are higher than their corresponding thresholds. The values for thresholds are sought so that the topology of the resulting network satisfies several criteria. The sought network has to be sparse, small-world (with any two genes being connected by a path of a few links only), scale-free-like (wherein a small number of genes have a large number of links and many genes have only a few connections). Similar to gene networks of other organisms the highly connected genes (hubs) in the constructed network tend to have essential cell functions. To verify the proposed method and to compare the results, a scale-free-like, small-world gene expression network was also constructed using another dataset [Le Roch, K.G., et al., *Science*, **301**(5639) (2003), 1503–1508], confirming the lethality and centrality property of malaria hubs.

Keywords: gene expression network, partial correlation, scale-free-like network

1. INTRODUCTION

The objective of this study is to construct a gene expression network of *Plasmodium falciparum* using the time-course microarray data-set from Bozdech et al. [3]. Unravelling the topology of the malaria gene network is relevant to understanding cell function and the invasion cycle of the parasite. We use a graph-theoretical approach where nodes in the network stand for genes and edges between two nodes stand for links representing relationships or associations between the two genes. In the network, the genes (nodes) are connected if certain criteria, such as co-expression, are satisfied.

Analyses of gene co-expression networks have shown a correlation between the essentiality of a gene and the number of connections that the gene has: highly connected genes (hubs) are often essential (involved in central biological functions) and evolutionarily conserved [2,16]. For *Plasmodium falciparum* more than 60% of predicted 5409 open reading frames lack sequence similarity to genes from any other known organism [3]. In addition, 65% of all annotated genes encode hypothetical proteins of unknown functions. This makes ascribing putative roles for such genes a challenging task. One of the potential benefits of gene network analysis is to obtain clues on the putative roles of such genes of unknown function based on the gene connectivities, positions in the network, and the other genes with which they have links.

It is of some interest to see whether the gene network analysis can give some support to the hypothesis advanced in [3] on a regulatory network wherein a comparatively small number of transcription factors with overlapping binding site specificities could account for the entire cascade. The authors speculated further that disruption of a key regulatory element (lethal gene) might have a profound inhibitory effect on the entire network [3]. Such lethal genes are most likely to be among the highly connected nodes in the malaria network.

For the study of the malaria gene regulatory network, we used two datasets. The first is the *overview* dataset from the complete intra-erythrocytic developmental cycle (IDC) transcriptome of *Plasmodium falciparum* measured at 46 time-points [3]. To verify results, we have also used a time-course dataset measured at nine time-points in human and mosquito stages of malaria parasite's life-cycle [10]. We will further refer to this dataset as the *validation* dataset.

2. NETWORK CONSTRUCTION FROM TOPOLOGICAL CONSTRAINTS

We aim to construct a network of malaria gene interactions, using global topology constraints, which have been found to be characteristic for other biological networks. These constraints include network sparseness, the small-world property, and the existence of a few highly connected nodes and many genes with a few connections.

An important measure of networks topology is the distribution of the number of connections per node. The number of connections per node is often called the *connectivity* of a node or its *degree*. Therefore, the distribution is referred to as the *connectivity (or degree) distribution*. Previously studied biological networks of interactions, including gene expression networks of other organisms, have shown to have many nodes with few connections and a few nodes with many connections (hubs) [1,2,11,16]. The existence of hubs has often been cited as the most characteristic feature of biological networks and in particular of the scale-free networks [1,2,16].

Although, the networks are commonly referred as being scale-free, it is their connectivity distribution that is considered to be scale-free. Precisely, distribution is defined as scale-free if its relative frequency distribution is given by a power-law, $p(k) \sim k^{-\gamma}$, $k \geqslant 1$, where k stands for the connectivity of a node and γ is the power-law exponent. It has recently been reported that the evidence collected to support the scale-free property of biological networks is questionable [15]. It has also been found that the connectivity distribution in many inferred biological networks differs in a statistically significant way from the power-law, and these networks are, strictly speaking, not scale-free [9]. In addition, a plausible evolutionary mechanism such as evolutionary drift is not compatible with scale-free distribution [13]. However, certain characteristics of a scale-free network, such as a small-world property and the existence of hubs, are valid for real genetic networks, and in the absence of consensus on an alternative distribution, the power-law can be used for modelling purposes as a first-order approximation. In particular, the connectivity distribution described by a power-law can be useful for construction of global gene regulatory networks, whose structure is mainly unknown. In this paper, we will be looking for the network connectivity distribution that resembles a power-law. We will refer to such networks as *scale-free-like*.

A chi-squared statistic $T = \sum_{k=1}^{k^*}(O_k - E_k)^2/E_k \sim \chi^2_{k^*-2}$ has been used as a measure of closeness of a network's connectivity distribution to scale-free behaviour. Here O_k are the observed (constructed) values of connectivities from the data, and E_k are the values estimated from the power-law, with γ estimated by the maximum likelihood method as described below. The connectivity values over k^*, for which the expected number of connections is less than 5, are pooled together. The smaller the value of the chi-squared statistic T the closer the connectivity distribution resembles a power-law.

For several gene co-expression networks, whose connectivity distribution has been modelled by the power-law, the power exponent γ has been reported to be of the order of 1.0 [2,16]. We have determined the power-exponent, $\hat{\gamma}$, of the network under consideration, by the maximum likelihood method from fitting the power-law distribution $p(k) = \frac{k^{-\gamma}}{\zeta(\gamma)}$ to the constructed connectivities (or degrees). Here $\zeta(\gamma)$ is the (truncated) Riemann zeta-function and $k \geqslant 1$. The number of connections (connectivity), x_i, for a node i is often obtained from experimental or simulated data.

In a large network the number of connections of different nodes can be assumed to be approximately independent. We have shown elsewhere [9] that the assumption of independence of connectivities of all nodes in the network can be weakened by assuming independence of connectivities of nodes in a smaller sub-network. As a result, the likelihood function can be written as $L(\gamma|x) = \prod_{i=1}^{N} x_i^{-\gamma}/\zeta(\gamma)$, where N is the maximum connectivity. The log-

likelihood $l(\gamma|x) = -\gamma \sum_{i=1}^{N} \log x_i - N \log \zeta(\gamma)$ is maximized by finding zeros of its derivative using the standard Newton–Raphson method for finding roots of the function.

Another important property gene expression networks have been shown to possess is a *small-world* property. In simple terms, this property implies that any two nodes can be connected with a path of only a few links. The small-world property is often quantitatively characterized by a large average clustering coefficient, C, which reflects the connectedness of the neighbours of a given node between themselves. The clustering coefficient of a gene i is computed by $c_i = 2n_i/k_i(k_i - 1)$, where n_i is the number of links connecting the k_i neighbours of gene i forming triangles and $k_i(k_i - 1)/2$ is the total number of triangles that could pass through the node i. The average clustering coefficient, C, of small-world networks is typically several orders of magnitude higher than that of a random network of equivalent average connectivity and size $C_r \approx k/N_{\text{genes}}$.

In addition, gene regulatory networks are known to be *sparse* because genes influence and/or are being influenced by a limited number of other genes [1]. This implies that average number of connections (connectivity) per gene (node), k is not large. Theoretical studies found the values for average connectivity in gene expression networks of different organisms to be of the order of 10–30 [11,16]. In this work, we will be looking for a network with the average connectivity in this range.

3. METHOD FOR THE CONSTRUCTION OF EXPRESSION NETWORK

The main thrust of this paper is to construct a malaria gene expression network based on thresholding pairwise Pearson correlations and partial correlations of gene profiles.

The threshold parameters were sought so that the constructed network satisfies four global topological criteria, described above. (1) The network is sparse, with an average connectivity, k, of the order of magnitude of 10; (2) the network has the small-world property such that is characterized by a clustering coefficient which is much higher than that of a random network with the same average connectivity and size, $C_r = 10/3000 = 0.003$; (3) the connectivity distribution is scale-free-like, i.e. it is as close as possible to the power-law, as seen in yeast and other organisms [2,4,11,16]; and (4) the power-law exponent $\hat{\gamma}$ of the connectivity distribution is close to 1.0 as has been reported for other gene expression networks [2,11,16].

3.1. Pearson Correlation

There have been a number of studies where global gene networks are constructed from microarray data based on the Pearson correlation coefficients. Two genes are considered linked in the co-expression network if their correlation is higher than the threshold [2,16]. Sometimes one also takes into account empirically calculated p-values for the correlations between two genes [4]. The Pearson correlation has been shown to play an important role in inferring interactions between genes [7]. However, methods that are based only on standard correlations are too simplistic and inevitably overestimate the number of links (connectivity) per gene. It is common knowledge that a high correlation coefficient is indicative not only of nodes that have direct connections but also of nodes with indirect connections. It is also plausible that some important true connections are left out if the threshold is not low enough. However, lowering the correlation threshold will significantly increase the number of potential links, including many random ones.

In the case of the malaria time-course dataset, the problem of including too many random links becomes even more transparent due to a very highly coordinated expression of genes [3]. A network constructed from the overview malaria dataset by thresholding correlations, while restricting the average connectivity per node, k, results in very high threshold values, R. For example, to obtain a network with $k = 50$ the threshold $R = 0.935$ is required. Restricting the average connectivity to a lower value, $k = 30$, results in an even higher value of threshold, $R = 0.95$. This is an unreasonably high value. Given the noisy data, missing values and the complexity of biological networks, many biologically relevant connections will not be included in such network.

For a slightly lower value of Pearson correlation cut-off, $R = 0.8$, the constructed network ceases to be sparse. In addition, its connectivity distribution is not scale-free-like (Figure 1). In fact, this co-expression network includes about 15% of all possible links, with an average number of links per node, $k = 470$, being more than ten times higher than the average connectivity for the gene networks of other organisms constructed by the same method. For example, with $k = 32$, the sparse scale-free network of yeast was constructed with only $R = 0.6$ [16].

3.2. Partial Correlation

Here we propose to use partial correlations to filter the more likely links out of a much larger set of potential links with high standard correlations. The partial correlation coefficient of two genes measures the strength of relation between these genes after the effect of other genes is removed or fixed, therefore indicating whether two genes are directly or indirectly linked. The partial correlations of different orders have been used in Gaussian Graphical

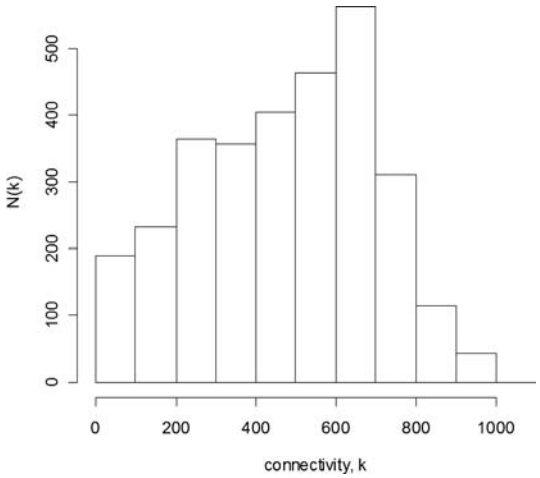

Figure 1. Histogram of connectivities in a malaria co-expression network constructed with a threshold $P = 0.8$ from overview dataset [3]. The average connectivity per node is $k = 470$ and the network is not scale-free. There are several highly connected genes and a much larger number of genes links with connectivities in the medium range.

Models (GGM) to characterize strength of correlations between pairs of genes in regulatory networks [12,14,17]. First-order partial correlations have been used to elucidate the regulatory network of *Arabidopsis thaliana* [17] and *Saccharomycces cerevisiae* [12]. These authors consider all possible triangles of three genes to explore the dependence between two of the genes conditioned on the third. All these triangles are then combined to make inferences on the complete network using either frequentist or latent random graph approaches. Second-order partial correlations, conditioning each pair of genes on every other pair of genes, have been applied to computer simulated networks and to yeast gene expression data [5]. Another method uses full-order partial correlations (conditioned on all other genes in the network) and the false discovery rate (FDR) approach to infer edges of the gene network from both simulated and real microarray data [14].

We propose to construct a gene expression network from a large gene dataset by using both Pearson and (full-order) partial correlation coefficients for each pair of genes. Namely, for each pair of genes (i, j) we compute the Pearson correlation of their profiles, r_{ij} , and their partial correlation coefficient, q_{ij} . The partial correlation of genes i and j with respect to other genes whose effect is removed (fixed) is given by

$$q_{ij} = \frac{\omega_{ij}}{\sqrt{\omega_{ii}\omega_{jj}}},$$

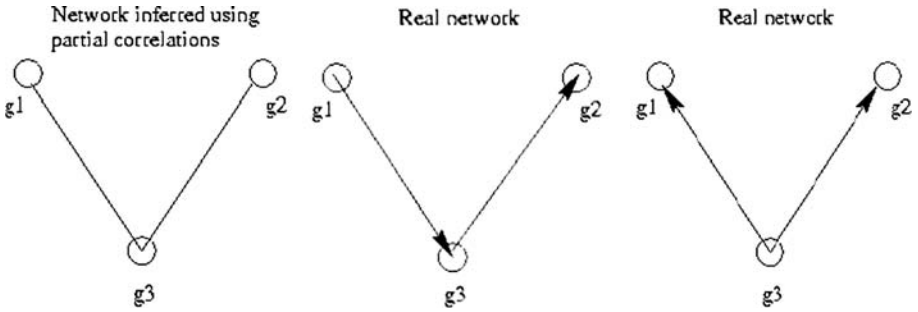

Figure 2. Schematic figure of the drawbacks of a representation of gene regulatory relationships by an undirected network. If in the inferred network, gene g3 is connected to genes g1 and g2 by undirected links (left), then it is impossible to distinguish between several scenarios in the real network. For example, gene g1 regulates gene g3, which in turn regulates gene g2 (middle), or gene g3 regulates genes g1 and g2 (right). Two other variants are possible.

where $\omega_{ij} = \{r_{ij}\}^{-1}$ is the inverse of the Pearson correlation matrix, $\{r_{ij}\}$. To overcome the degeneracy problem of the correlation matrix $\{r_{ij}\}$ for small samples, partial correlation estimators based on the Moore–Penrose pseudo-inverse of correlation matrix were introduced in [14]. In our work we follow this approach and compute partial correlations by using the Moore–Penrose pseudo-inverse of the correlation matrix via the *cor2pcor*( ) function from *R*-package *GeneTS* [14]. Two genes (i, j) are connected by a link if their Pearson correlation is higher than a cut-off value, R, and their partial correlation is higher than (or equal to) a cut-off value, Q:

$$i \leftrightarrow j\colon\ r_{ij} \geqslant R \text{ and } q_{ij} \geqslant Q.$$

The general drawback of any inference approach that results in an undirected network (such as a GGM) is that it gives no indication of causality. A link connecting two genes does not indicate which gene in the pair is the regulator and which is the regulated one, as illustrated in Figure 2. Although lacking causality information, undirected networks are a very useful first level representation of gene regulatory relationships on a genome wide level. Further levels of representations are directed networks, where the direction of the regulatory relationship is specified. This can eventually be extended by quantitative information, such as probabilities of connection in Bayesian networks or kinetic parameters of regulation.

4. RESULTS

For the overview dataset, the values from multiple oligonucleotides representing the same gene were averaged, resulting in expression values for 3048

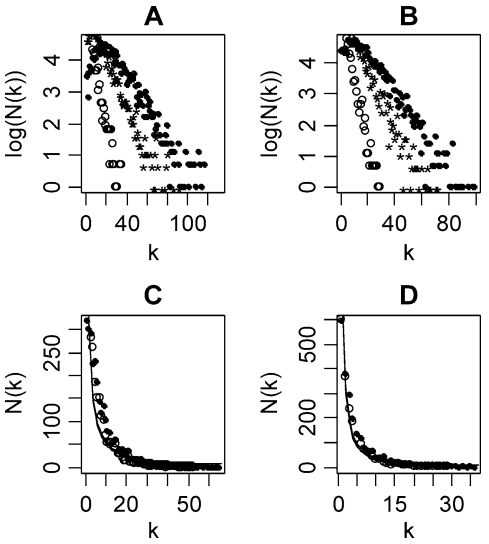

Figure 3. Connectivity distribution of nodes in a malaria gene network constructed from the overview dataset for different values of thresholds. (A), (B) Log distribution of connectivities for $r = 0.45; 0.5; 0.55$ and $P = 0.7$ (A) and $P = 0.8$ (B). (C), (D) Distribution of observed connectivities and fitted power-law $N(k) \sim k^{-\gamma}$ for $r = 0.5$ and $P = 0.8$, $\hat{\gamma} = 0.91$ (C) and $P = 0.7$, $\hat{\gamma} = 0.84$ (D).

genes. In the rest of the paper we will concentrate on reporting the results for the overview dataset. The topology of the network constructed using the validation dataset is very similar (see Tables S5 and S6 on the supplemental web-page: www.stats.gla.ac.uk/~raya/Malaria/suppldata.html).

We have performed a grid search for the threshold values R and Q based on topological criteria. We have found a range $0.45 \leqslant Q \leqslant 0.6$ and $0.7 \leqslant R \leqslant 0.8$ for which all four topological constraints are satisfied. The qualitative topological properties of the malaria network are insensitive to the precise thresholds within this range of values. Taking the thresholds within this range yields a scale-free-like distribution of connectivities, which are qualitatively similar. Figure 3 shows connectivity distributions, $N(k)$, for several values of thresholds R and Q. Values outside this region result in other types of networks. $Q \leqslant 0.4$ results in networks whose connectivities do not obey a power-law (Figure 4); while $Q > 0.6$ and/or $R > 0.8$ yield too few links (not shown).

Values of $\hat{\gamma}$ are within the range 0.6–1.4 for different values of thresholds Q and R. $\gamma = 0.6$ for the parameters $R = 0.7$, $Q = 0.45$ produce a network with an average connectivity per node of $k = 28$ and maximum connectivity $k_{\max} = 133$, and $\hat{\gamma} = 1.4$ is for $R = 0.8$, $Q = 0.6$ with $k = 4$, $k_{\max} = 30$. Other values of parameters resulted in networks with average connectivities between these two values (see Table S1 on the supplemental web-page).

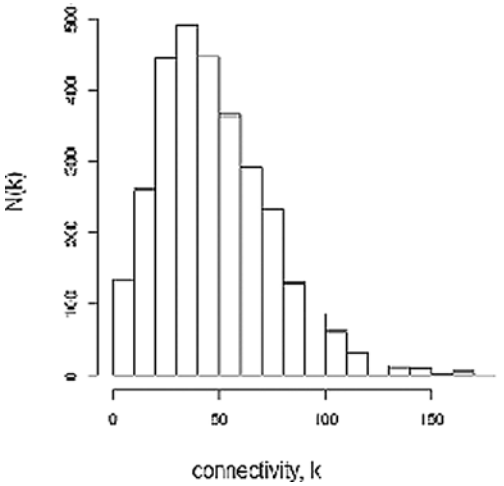

Figure 4. Histogram of connectivities in a malaria co-expression network constructed with thresholds $P = 0.7$, $r = 0.4$ from overview dataset [3]. Lowering one of the thresholds outside the accepted region results in a network whose behaviour is very different from scale-free.

The clustering coefficients have been found to be within the range $C = 0.195$ for $R = 0.7$, $Q = 0.45$ and $C = 0.443$ for $R = 0.8$, $Q = 0.6$. These values are much higher than the value for random networks of equivalent average connectivity and size ($C = 0.003$), and they are consistent with the values reported for other organisms (e.g., $C = 0.6$ for yeast [16] and other organisms [2]).

4.1. Statistical Validation

To find whether a network constructed by thresholding the two types of correlation coefficients is statistically meaningful or whether it can easily be found by chance, we performed a permutation test. For each gene we reshuffle the values at each time-point, constructing a gene profile of the same length, with the same values but with a different time-order of these values. We then recompute the correlation and partial correlation matrices and establish a link between genes i and j if the thresholding conditions ($R = 0.7$, $Q = 0.5$) are satisfied. In 100 permutation networks, two links are found on average for each network (estimated standard error $= 0.14$) compared to several thousands in the network inferred from the original dataset. This allows us to conclude that the network inferred by the thresholding method is unlikely to have arisen by chance.

4.2. Biological Validation

The expression network constructed by the proposed method in Section 4.1 is worth investigating further for some proof-of-principle results. In the next section we report results for the threshold values $R = 0.7$, $Q = 0.5$. These parameters yield network statistics that are similar to previously studied networks with a maximum connectivity $k_{\max} = 101$, average connectivity per node $k = 15$, and the power-law exponent $\hat{\gamma} = 0.84$.

4.2.1. Lethality and Centrality of Malaria Genes. It has been previously reported that high degree nodes in gene expression networks constructed for other organisms are more likely to correspond to essential and conserved genes, i.e. to be involved in central biological functions of the cell [2]. In the constructed network, among the top 66 hubs with connectivities from $k_{\max}/2$ there are 13 genes with no manual annotation, 7 genes belong to the Plastid genome, and 30 genes code for proteins with unknown functions (hypothetical proteins). Therefore, only 16 hub-genes code for proteins with some identifiable functions. Among them, 7 genes (PFI1340w, PFI1360c, PFI0385c, PF13_0229, PF14_0373, PFA0345w, PF11_0298) are known to have essential functions in cell growth, maintenance, and metabolism (according to GO annotation). In addition, a rhoptry protein (PFI0265c), a papain family cysteine protease (PFI0135c), and an early transcribed membrane protein (PF10_0019) are also in the list of the hub-genes. Among 5 hubs on chromosome 9, three (PFI1340w, PFI1360c, and PFI0385c) are prescribed functions in cell growth, maintenance and metabolism, and they are all connected among themselves forming a triangular network motif. The largest reported ORF (MAL6P1.147) also has a large number of links, half of maximum connectivity. Other 8 annotated hubs out of 30 that code for hypothetical proteins are either conserved or have homologues/similar to proteins in other organisms.

The list of 66 top hubs for the network constructed from the validation dataset with $R = 0.8$, $Q = 0.5$ contains 20 genes (virtually all annotated hubs) with cell growth/maintenance, cell communication, and other central cell functions. For a full list of hubs in networks constructed for the overview and the validation datasets see Tables S2 and S6 on the supplemental webpage.

As another proof-of-principle, we looked at how many hubs are in the set of only 6% of all genes in the genome of *Plasmodium falciparum* that were found to be common to all four stages of the parasite life cycle (supplementary Table 1 in [6]). This list contains primarily housekeeping genes and their products, such as ribosomal proteins, transcription factors, and cytoskeletal proteins. It turned out that 15 hubs from our list are among this set of common genes found in [6]. This is about 30% of all hubs with manual annotation.

It is of interest to see whether genes with unknown functionality among the hubs belong to classes of essential genes. We looked at how hubs that code

for protein with unknown functions in the overview network clustered in the experiments of Le Roch et al. [10], as it has been demonstrated for various organisms that genes that cluster together are more likely to have similar biological functions. We found that among 25 genes coding for hypothetical proteins that are present in the validation dataset, 10 genes belong to cluster 13, 5 to cluster 12, and 5 to cluster 15 of [10]. Le Roch et al. [10] reported that genes of known functions in clusters 12, 13 are mainly involved in cell-cycle regulation and progression at trophozoite stage, while cluster 15 is characterized as having genes with roles in cell invasion that are under evaluation as blood-stage vaccine. Therefore, hubs of unknown functions in those clusters are more likely to be of these essential functions. It is worth mentioning that, according to the authors of [10], genes from clusters 12 and 13 may represent potential targets for drugs focused on disruption of the highly replicating trophozoite stage of the parasite, while additional candidate vaccine antigens could come from the yet uncharacterized genes of cluster 15. This gives further support to our conjecture that the hubs of unknown functions might be of important biological functions and therefore warrant further investigation.

We believe that the above mentioned evidence demonstrates that the hubs in the constructed malaria gene network tend to be essential. It will also be interesting to investigate those genes among hubs that have not been manually annotated (see Table S3 that contains oligonucleotides of hubs from the overview network with no manual annotation).

4.2.2. Some Sub-networks in Malaria Gene Network. It might be interesting to investigate further some sub-networks of the large malaria gene network. As an example, we had a closer look at the glycolytic pathway, as it is mentioned in [3] as the one that is well-preserved in malaria parasite. Among 9 genes from the microarray dataset that belong to this pathway as taken from the http://plasmodb.org database, we found that they share 5 links among themselves. In fact, the probability of 9 randomly picked genes to have 5 links is 0.01% given the connectivity matrix. Given that some of the genes in this pathway are not present in the dataset, this result is encouraging. Our analysis did not pick up MAL61.160 as part of the glycolytic pathway. Instead, another putative copy, PF10_0363, was identified as a part of it, having 2 connections, as well as gene PF10_0155 that has 4 connections.

As another example, we had a look at all major candidates for vaccination (AMA1, EBA175, MSP1, MSP3, MSP7, RAP1, RESA1) studied in [3]. All these genes are very well positioned in the network, having connectivities between 20 and 40, well above the average connectivity of $k = 15$. Interestingly, these vaccine candidates are connected among themselves as well as with some other merozoite invasion proteins (MSP6, MSP8). In addition, the neighbours

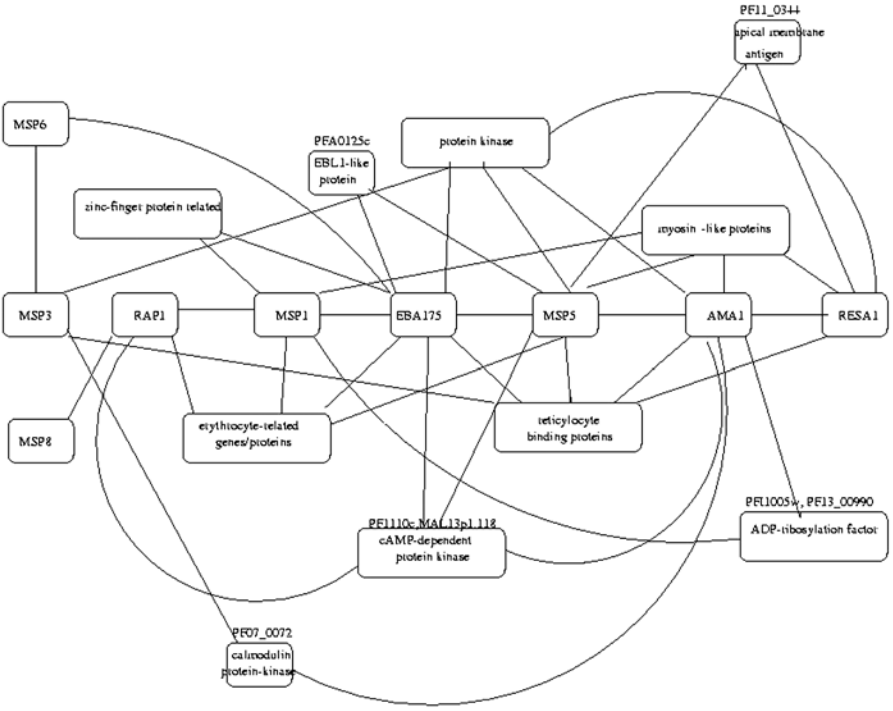

Figure 5. Sketch of a sub-network of seven major malaria vaccine candidates. The sub-network contains major vaccine candidates (AMA1, EBA175, MSP1, MSP3, MSP7, RAP1, RESA1) that have been studied in [3] and some genes/proteins or groups according to our model. Small boxes contain one gene/protein and larger boxes contain two or more related genes/proteins. Only some links are shown. For a full list of links of the seven major malaria vaccine candidates according to our model see Tables S4 on supplemental page.

of these vaccine candidates are enriched with myosin-like proteins, erythrocyte associated proteins, reticulocyte binding proteins, and zinc finger proteins among others (Figure 5). For example, the erythrocyte-related group contains erythrocyte binding antigenes (PF07_0128, MAL13P1.60), erythrocyte surface antigene (PFA0110w), and erythrocyte binding proteins (PF08_0142, PF08_0147). The myosin-like proteins group contains 4 genes (PF13_0233, PFL225w, MAL6P1.286, PFL1435c). There are 4 genes in the reticulocyte-binding proteins group (PF13_0198, PFL2520w, MAL13P1.176, PFD0100w). The protein-kinase group includes PF130815w, PFC0945w, PFB00150c, and Ser/Thr protein-kinase PFB0665w. The zinc-finger related group contains one zinc-finger protein (PFE0895) and a cell-cycle regulator with zinc-finger domain (PFE1415w). There are a large number of hypothetical proteins that are linked to the vaccine candidates in our network. Several of the hypothetical proteins from the list are linked to two major vaccine candidates, while some

hypothetical proteins (e.g., PF10_0352, PF07_0127, PFE0365c, PFC1045c, PFD0715c) have links with three major vaccine candidates and are probably worth having a closer look at. For a full list of the neighbours of these major vaccine candidates see Table S4.

5. CONCLUSIONS

In this paper we have constructed a model of malaria gene expression network by a novel method of thresholding two types of pair-wise correlation coefficients: the Pearson correlation and the full-order partial correlation coefficients. The values for thresholds were determined by topological considerations. Both types of correlations are essential in revealing the connections of genes in the network. The constructed small-world, scale-free network has hub-genes that tend to have essential cell functions, similar to other biological networks. We propose that hubs with unknown functions warrant further investigation in the search for malaria vaccine. Finding hubs in the malaria gene network is extremely important in guiding the search for the malaria vaccine. Targeting a highly connecting node with a drug will result in inactivation of a protein that could be fatal to the whole life-cycle of the malaria parasite, whereas removing a less connected node will barely affect the whole system.

This model of malaria gene network is worth investigating further by looking at various sub-networks consisting of genes that are known to be involved in the same biological processes. Alternatively, one might want to look at the neighbours of genes with unknown functions. This might help the process of assigning putative functions to these genes. The links adjacency matrix of the network studied in this paper can be found on the supplemental web-page: www.stats.gla.ac.uk/~raya/Malaria/suppldata.html. To summarise, the thresholding approach of two correlation coefficients that is proposed in this paper suffices for the goal of studying statistical properties of a biological network and also gives encouraging proof-of-principle results.

REFERENCES

[1] Barabasi, A.L. and Oltvai, Z.N., Network biology: Understanding the cell's functional organization, *Nat. Rev. Genet.*, **2** (2004), 101–113.

[2] Bergmann, S., Ihmels, J., and Barkai, N., Similarities and differences in genome-wide expression data of six organisms, *PLoS Biol.*, **2**(1) (2004), E9.

[3] Bozdech, Z., Llinas, M., Pulliam, B.L., Wong, E.D., Zhu, J., and DeRisi, J.L., The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*, *PLoS Biol.*, **1**(1) (2003), E5.

[4] Carter, S.L., Brechbuhler, C.M., Griffin, M., and Bond, A.T., Gene expression network topology provides a framework for molecular characterization of cellular state, *Bioinformatics*, **20** (2004), 2242–2250.

[5] de la Fuente, A., Bing, N., Hoeschele, I., and Mendes, P., Discovery of meaningful associations in genomic data using partial correlation coefficients, *Bioinformatics*, **20** (2004), 3565–3574.

[6] Florens, L., Washburn, M.P., Raine, J.D., Anthony, R.M., Grainger, M., Haynes, J.D., Moch, J.K., Muster, N., Sacci, J.B., Tabb, D.L., Witney, A.A., Wolters, D., Wu, Y., Gardner, M.J., Holder, A.A., Sinden, R.E., Yates, J.R., and Carucci, D.J., A proteomic view of the *Plasmodium falciparum* life cycle, *Nature*, **3**(419) (2002), 520–526.

[7] Ideker, T., Thorsson, V., Ranish, J.A., Christmas, R., Buhler, J., Eng, J.K., Bumgarner, R., Goodlett, D.R., Aebersold, R., and Hood, L., Integrated genomic and proteomic analyses of a systematically perturbed metabolic network, *Science*, **929**(5518) (2001), 929–934.

[8] Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N., and Barabasi, A.L., The large-scale organization of metabolic networks, *Nature*, **407**(6804) (2000), 651–654.

[9] Khanin, R. and Wit, E., How scale-free are biological networks, (2005), submitted. www.stats.gla.ac.uk/~raya/howscalefree/howscalefree.pdf

[10] Le Roch, K.G., Zhou, Y., Blair, P.L., Grainger, M., Moch, J.K., Haynes, J.D., De La Vega, P., Holder, A.A., Batalov, S., Carucci, D.J., and Winzeler, E.A., Discovery of gene function by expression profiling of the malaria parasite life cycle, *Science*, **301**(5639) (2003), 1503–1508.

[11] Luscombe, N.M., Babu, M.M., Yu, H., Snyder, M., Teichmann, S.A., and Gerstein, M., Genomic analysis of regulatory network dynamics reveals large topological changes, *Nature*, **431** (2004), 308–312.

[12] Magwene, P.M. and Kim, J., Estimating genomic coexpression networks using first-order conditional independence, *Genome Biology*, **5** (2004), R100.

[13] Przytycka, T. and Yu, Y.-K., Scale-free networks versus evolutionary drift, *Comput. Biol. Chem.*, **28** (2004), 257–264.

[14] Schafer, J. and Strimmer, K., An empirical Bayes approach to inferring large graphical Gaussian models from microarray data, *Bioinformatics*, in press. http://www.stat.uni-muenchen.de/~strimmer/publications/largeggm2004.pdf

[15] Stumpf, M.P., Wiuf, C., and May, R.M., Subnets of scale-free networks are not scale-free: Sampling properties of networks, *Proc. Natl. Acad. Sci. USA*, **102**(12) (2005), 4221–4224.

[16] van Noort, V., Snel, B., and Huynen, M.A., The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model, *EMBO Rep.*, **5**(3) (2004), 280–284.

[17] Wille, A., Zimmermann, P., Vranova, E., Furholz, A., Laule, O., Bleuler, S., Hennig, L., Prelic, A., von Rohr, P., Thiele, L., Zitzler, E., Gruissem, W., and Buhlmann, P., Sparse graphical modeling of the iroprenoid gene network in *Arabidopsis thaliana*, *Genome Biology*, **5** (2004), 11.